POSSIBLE MINDS

25 WAYS OF LOOKING AT AI

EDITED BY

JOHN BROCKMAN

Seth Lloyd
Judea Pearl
Stuart Russell
George Dyson
Daniel C. Dennett
Rodney Brooks
Max Tegmark
Venki Ramakrishnan

Frank Wilczek
Jaan Tallinn
Steven Pinker
David Deutsch
Tom Griffiths
Anca Dragan
Chris Anderson
David Kaiser
Neil Gershenfeld

W. Daniel Hillis
Hans Ulrich Obrist
Alison Gopnik
George M. Church
Caroline A. Jones
Alex “Sandy” Pentland
Stephen Wolfram
Peter Galison

BOOKS BY JOHN BROCKMAN

AS AUTHOR

By the Late John Brockman
37
Afterwords
The Third Culture
Digerati

AS EDITOR

About Bateson
Speculations
Doing Science
Ways of Knowing
Creativity
The Greatest Inventions of the Past 2,000 Years
The Next Fifty Years
The New Humanists
Curious Minds
What We Believe but Cannot Prove
My Einstein
Intelligent Thought
What Is Your Dangerous Idea?
What Are You Optimistic About?
Science at the Edge
What Have You Changed Your Mind About?
This Will Change Everything
Is the Internet Changing the Way You Think?
Culture
The Mind
This Will Make You Smarter
This Explains Everything
Thinking
What Should We Be Worried About?
The Universe
This Idea Must Die
What to Think About Machines That Think
Life
Know This
This Idea Is Brilliant

AS CO-EDITOR

How Things Are (with Katinka Matson)

POSSIBLE MINDS

Twenty-Five Ways of Looking at AI

Edited by

JOHN BROCKMAN

PENGUIN PRESS ‖ NEW YORK ‖ 2019

PENGUIN PRESS
An imprint of Penguin Random House LLC
penguinrandomhouse.com

LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA

Names: Brockman, John, 1941- editor.
Title: Possible minds : twenty-five ways of looking at AI / edited by John Brockman.
Description: New York : Penguin Press, 2019. | Includes index.
Identifiers: LCCN 2018032888 | ISBN 9780525557999 (hardcover) | ISBN 9780525558002 (ebook)
Subjects: LCSH: Artificial intelligence—Social aspects.
Classification: LCC Q335 .D436 2018 | DDC 006.3—dc23
LC record available at https://lccn.loc.gov/2018032888

Version_1

For Einstein, Gertrude Stein, Wittgenstein, and Frankenstein

Acknowledgments

My thanks to Scott Moyers of Penguin Press for his editorial exuberance and my agent, Max Brockman, for his continued encouragement. A special thanks, once again, to Sara Lippincott for her thoughtful attention to the manuscript.

CONTENTS

The most dramatic discontinuity will surely be when we achieve effective human immortality. Whether this will be achieved biologically or digitally isn't clear, but inevitably it will be achieved.

INTRODUCTION: ON THE PROMISE AND PERIL OF AI

Artificial intelligence is today's story—the story behind all other stories. It is the Second Coming and the Apocalypse at the same time: good AI versus evil AI. This book comes out of an ongoing conversation with a number of important thinkers, both in the world of AI and beyond it, about what AI is and what it means. Called the Possible Minds Project, this conversation began in earnest in September 2016, in a meeting at the Grace Mayflower Inn & Spa in Washington, Connecticut, with some of the book's contributors.

What quickly emerged from that first meeting is that the excitement and fear in the wider culture surrounding AI now has an analog in the way Norbert Wiener's ideas regarding "cybernetics" worked their way through the culture, particularly in the 1960s, as artists began to incorporate thinking about new technologies into their work. I witnessed the impact of those ideas at close hand; indeed, it's not too much to say they set me off on my life's path. With the advent of the digital era beginning in the early 1970s, people stopped talking about Wiener, but today, his Cybernetic Idea has been so widely adopted that it's internalized to the point where it no longer needs a name. It's everywhere, it's in the air, and it's a fitting place to begin.

NEW TECHNOLOGIES = NEW PERCEPTIONS

Before AI, there was cybernetics—the idea of automatic, self-regulating control, laid out in Norbert Wiener's foundational text of 1948. I can date my own serious exposure to it to 1966, when the composer John Cage invited me and four or five other young arts people to join him for a series of dinners—an ongoing seminar about media, communications, art, music, and philosophy that focused on Cage's interest in the ideas of Wiener, Claude Shannon, and Marshall McLuhan, all of whom had currency in the New York art circles in which I was then moving. In particular, Cage had picked up on McLuhan's idea that by inventing electronic technologies we had externalized our central nervous system—that is, our minds—and that we now had to presume that "there's only one mind, the one we all share."

Ideas of this nature were beginning to be of great interest to the artists I was working with in New York at the Film-Makers' Cinematheque, where I was program manager for a series of multimedia productions called the New Cinema 1 (also known as the Expanded Cinema Festival), under the auspices of avant-garde filmmaker and impresario Jonas Mekas. They included visual artists Claes Oldenburg, Robert Rauschenberg, Andy Warhol, and Robert Whitman; kinetic artists Charlotte Moorman and Nam June Paik; happenings artists Allan Kaprow and Carolee Schneemann; dancer Trisha Brown; filmmakers Jack Smith, Stan Vanderbeek, Ed Emshwiller, and the Kuchar brothers; avant-garde dramatist Ken Dewey; poet Gerd Stern and the USCO group; minimalist musicians La Monte Young and Terry Riley; and, through Warhol, the music group The Velvet Underground. Many of these people were reading Wiener, and cybernetics was in the air. It was at one of these dinners that Cage reached into his briefcase and took out a copy of *Cybernetics* and handed it to me, saying, "This is for you."

During the festival, I received an unexpected phone call from Wiener's colleague Arthur K. Solomon, head of Harvard's graduate program in biophysics. Wiener had died the year before, and Solomon's and Wiener's other close colleagues at MIT and Harvard had been reading about the Expanded Cinema Festival in the *New York Times* and were intrigued by the connection to Wiener's work. Solomon invited me to bring some of the artists up to Cambridge to meet with him and a group that included MIT sensory-communications researcher Walter Rosenblith, Harvard

applied mathematician Anthony Oettinger, and MIT engineer Harold "Doc" Edgerton, inventor of the strobe light.

Like many other "art meets science" situations I've been involved in since, the two-day event was an informed failure: ships passing in the night. But I took it all on board and the event was consequential in some interesting ways—one of which came from the fact that they took us to see "the" computer. Computers were a rarity back then; at least none of us on the visit had ever seen one. We were ushered into a large space on the MIT campus, in the middle of which there was a "cold room" raised off the floor and enclosed in glass, in which technicians wearing white lab coats, scarves, and gloves were busy collating punch cards coming through an enormous machine. When I approached, the steam from my breath fogged up the window into the cold room. Wiping it off, I saw "the" computer. I fell in love.

Later, in the fall of 1967, I went to Menlo Park to spend time with Stewart Brand, whom I had met in New York in 1965 when he was a satellite member of the USCO group of artists. Now, with his wife, Lois, a mathematician, he was preparing the first edition of the *Whole Earth Catalog* for publication. While Lois and the team did the heavy lifting on the final mechanicals for *WEC*, Stewart and I sat together in a corner for two days, reading, underlining, and annotating the same paperback copy of *Cybernetics* that Cage had handed to me the year before, and debating Wiener's ideas.

Inspired by this set of ideas, I began to develop a theme, a mantra of sorts, that has informed my endeavors since: "new technologies = new perceptions." Inspired by communications theorist Marshall McLuhan, architect-designer Buckminster Fuller, futurist John McHale, and cultural anthropologists Edward T. "Ned" Hall and Edmund Carpenter, I started reading avidly in the fields of information theory, cybernetics, and systems theory. McLuhan suggested I read biologist J. Z. Young's *Doubt and Certainty in Science,* in which he said that we create tools and we mold ourselves through our use of them. The other text he recommended was Warren Weaver and Claude Shannon's 1949 paper "Recent Contributions to the Mathematical Theory of Communication," which begins: "The word *communication* will be used here in a very broad sense to include all of the procedures by which one mind may affect another.

This, of course, involves not only written and oral speech, but also music, the pictorial arts, the theater, the ballet, and in fact all human behavior."

Who knew that within two decades of that moment we would begin to recognize the brain as a computer? And in the next two decades, as we built our computers into the Internet, that we would begin to realize that the brain is not a computer but a network of computers? Certainly not Wiener, a specialist in analog feedback circuits designed to control machines, nor the artists, nor, least of all, myself.

"WE MUST CEASE TO KISS THE WHIP THAT LASHES US."

Two years after *Cybernetics,* in 1950, Norbert Wiener published *The Human Use of Human Beings*—a deeper story, in which he expressed his concerns about the runaway commercial exploitation and other unforeseen consequences of the new technologies of control. I didn't read *The Human Use of Human Beings* until the spring of 2016, when I picked up my copy, a first edition, which was sitting in my library next to *Cybernetics*. What shocked me was the realization of just how prescient Wiener was in 1950 about what's going on today. Although the first edition was a major bestseller—and, indeed, jump-started an important conversation—under pressure from his peers Wiener brought out a revised and milder edition in 1954, from which the original concluding chapter, "Voices of Rigidity," is conspicuously absent.

Science historian George Dyson points out that in this long-forgotten first edition, Wiener predicted the possibility of a "threatening new Fascism dependent on the *machine à gouverner*":

> No elite escaped his criticism, from the Marxists and the Jesuits ("all of Catholicism is indeed essentially a totalitarian religion") to the FBI ("our great merchant princes have looked upon the propaganda technique of the Russians, and have found that it is good") and the financiers lending their support "to make American capitalism and the fifth freedom of the businessman supreme

throughout the world." Scientists . . . received the same scrutiny given the Church: "Indeed, the heads of great laboratories are very much like Bishops, with their association with the powerful in all walks of life, and the dangers they incur of the carnal sins of pride and of lust for power."

This jeremiad did not go well for Wiener. As Dyson puts it:

> These alarms were discounted at the time, not because Wiener was wrong about digital computing but because larger threats were looming as he completed his manuscript in the fall of 1949. Wiener had nothing against digital computing but was strongly opposed to nuclear weapons and refused to join those who were building digital computers to move forward on the thousand-times-more-powerful hydrogen bomb.

Since the original of *The Human Use of Human Beings* is now out of print, lost to us is Wiener's cri de coeur, more relevant today than when he wrote it sixty-eight years ago: "We must cease to kiss the whip that lashes us."

MIND, THINKING, INTELLIGENCE

Among the reasons we don't hear much about cybernetics today, two are central: First, although *The Human Use of Human Beings* was considered an important book in its time, it ran counter to the aspirations of many of Wiener's colleagues, including John von Neumann and Claude Shannon, who were interested in the commercialization of the new technologies. Second, computer pioneer John McCarthy disliked Wiener and refused to use Wiener's term "Cybernetics." McCarthy, in turn, coined the term "artificial intelligence" and became a founding father of that field.

As Judea Pearl, who, in the 1980s, introduced a new approach to artificial intelligence called Bayesian networks, explained to me:

> What Wiener created was excitement to believe that one day we are going to make an intelligent machine. He wasn't a computer scientist. He talked feedback, he talked communication, he talked analog. His working metaphor was a feedback circuit, which he was an expert in. By the time the digital age began in the early 1960s people wanted to talk programming, talk codes, talk about computational functions, talk about short-term memory, long-term memory—meaningful computer metaphors. Wiener wasn't part of that, and he didn't reach the new generation that germinated with his ideas. His metaphors were too old, passé. There were new means already available that were ready to capture the human imagination. By 1970, people were no longer talking about Wiener.

One critical factor missing in Wiener's vision was the cognitive element: mind, thinking, intelligence. As early as 1942, at the first of a series of foundational interdisciplinary meetings about the control of complex systems that would come to be known as the Macy Conferences, leading researchers were arguing for the inclusion of the cognitive element into the conversation. While von Neumann, Shannon, and Wiener were concerned about systems of control and communication of observed systems, Warren McCullough wanted to include mind. He turned to cultural anthropologists Gregory Bateson and Margaret Mead to make the connection to the social sciences. Bateson, in particular, was increasingly talking about patterns and processes, or "the pattern that connects." He called for a new kind of systems ecology in which organisms and the environment in which they live are one and the same and should be considered as a single circuit. By the early 1970s the cybernetics of observed systems—first-order cybernetics—moved to the cybernetics of observing systems—second-order cybernetics, or "the Cybernetics of Cybernetics," as coined by Heinz von Foerster, who joined the Macy Conferences in the mid-1950s and spearheaded the new movement.

Cybernetics, rather than disappearing, was becoming metabolized into *everything*, so we no longer saw it as a separate, distinct new discipline. And there it remains, hiding in plain sight.

“THE SHTICK OF THE STEINS”

My own writing about these issues at the time was on the radar screen of the second-order cybernetics crowd, including Heinz von Foerster as well as John Lilly and Alan Watts, who were the co-organizers of something called the AUM Conference, shorthand for “the American University of Masters,” which took place in Big Sur in 1973, a gathering of philosophers, psychologists, and scientists, each of whom was asked to lecture on his own work in terms of its relationship to the ideas of British mathematician G. Spencer-Brown as presented in his book *Laws of Form*. I was a bit puzzled when I received an invitation—a very late invitation indeed—which they explained was based on their interest in the ideas I presented in a book called *Afterwords*, which were very much on their wavelength. I jumped at the opportunity, the main reason being that the keynote speaker was none other than Richard Feynman. I love to spend time with physicists, because they think about the universe, i.e., everything. And no physicist was reputed to be as articulate as Feynman. I couldn’t wait to meet him. I accepted. That said, I am not a scientist, and I had never entertained the idea of getting on a stage and delivering a “lecture” of any kind, least of all a commentary on an obscure mathematical theory in front of a group identified as the world’s most interesting thinkers. Only upon my arrival in Big Sur did I find out the reason for my very late invitation. “When is Feynman’s talk?” I asked at the desk. “Oh, didn’t Alan Watts tell you? Richard is ill and has been hospitalized. You’re his replacement. And, by the way, what’s the title of your keynote lecture?”

I tried to make myself invisible for several days. Alan Watts, realizing that I was avoiding the podium, woke me up one night with a three a.m. knock on the door of my room. I opened the door to find him standing in front of me wearing a monk’s robe with a hood covering much of his face. His arms extended, he held a lantern in one hand and a magnum of scotch in the other. “John,” he said in a deep voice with a rich aristocratic British accent, “you are a phony. And, John,” he continued, “I am a phony. But, John, I am a *real* phony!”

The next day I gave my lecture, titled “Einstein, Gertrude Stein, Wittgenstein, and Frankenstein.” Einstein: the revolution in 20th century

physics. Gertrude Stein: the first writer who made integral to her work the idea of an indeterminate and discontinuous universe. Words represented neither character nor activity: A rose is a rose is a rose, and a universe is a universe is a universe. Wittgenstein: the world as limits of language. “The limits of my language mean the limits of my world.” The end of the distinction between observer and observed. Frankenstein: cybernetics, AI, robotics, all the essayists in this volume.

The lecture had unanticipated consequences. Among the participants at the AUM Conference were several authors of number one *New York Times* bestsellers, yet no one there had a literary agent. And I realized that all were engaged in writing a genre of book both unnamed and unrecognized by New York publishers. Since I had an MBA from Columbia Business School and a series of relative successes in business, I was dragooned into becoming an agent, initially for Gregory Bateson and John Lilly, whose books I sold quickly, and for sums that caught my attention, thus kick-starting my career as a literary agent.

I never did meet Richard Feynman.

THE LONG AI WINTERS

This new career put me in close touch with most of the AI pioneers, and over the decades I rode with them on waves of enthusiasm, and into valleys of disappointment. In the early eighties the Japanese government mounted a national effort to advance AI. They called it the Fifth Generation; their goal was to change the architecture of computation by breaking “the von Neumann bottleneck” by creating a massively parallel computer. In so doing, they hoped to jump-start their economy and become a dominant world power in the field. In 1983, the leader of the Japanese Fifth Generation consortium came to New York for a meeting organized by Heinz Pagels, the president of the New York Academy of Sciences. I had a seat at the table alongside the leaders of the first generation, Marvin Minsky and John McCarthy; the second generation, Edward Feigenbaum and Roger Schank; and Joseph Traub, head of the National Supercomputer Consortium.

In 1981, with Heinz's help, I had founded The Reality Club (the precursor to the nonprofit Edge.org), whose initial interdisciplinary meetings took place in the boardroom at the NYAS. Heinz was working on his book *The Dreams of Reason: The Computer and the Rise of the Science of Complexity*, which he considered to be a research agenda for science in the 1990s.

Through the Reality Club meetings, I got to know two young researchers who were about to play key roles in revolutionizing computer science. At MIT in the late seventies, Danny Hillis developed the algorithms that made possible the massively parallel computer. In 1983, his company, Thinking Machines, built the world's fastest supercomputer by utilizing parallel architecture. His "connection machine" closely reflected the workings of the human mind. Seth Lloyd at Rockefeller University was undertaking seminal work in the fields of quantum computation and quantum communications, including proposing the first technologically feasible design for a quantum computer.

And the Japanese? Their foray into artificial intelligence failed and was followed by twenty years of anemic economic growth. But the leading U.S. scientists took this program very seriously. And Feigenbaum, who was the cutting-edge computer scientist of the day, teamed up with Pamela McCorduck to write a book on these developments. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World* was published in 1983. We had a code name for the project: "It's coming, it's coming!" But it didn't come; it went.

From that point on I've worked with researchers in nearly every variety of AI and complexity, including Rodney Brooks, Hans Moravec, John Archibald Wheeler, Benoit Mandelbrot, John Henry Holland, Danny Hillis, Freeman Dyson, Chris Langton, J. Doyne Farmer, Geoffrey West, Stuart Russell, and Judea Pearl.

AN ONGOING DYNAMICAL EMERGENT SYSTEM

From the initial meeting in Washington, Connecticut, to the present, I arranged a number of dinners and discussions in London and Cambridge, Massachusetts, as well as a public event at London's City Hall. Among the attendees were distinguished scientists, science historians, and communications theorists, all of whom have been thinking seriously about AI issues for their entire careers.

I commissioned essays from a wide range of contributors, with or without references to Wiener (leaving it up to each participant). In the end, twenty-five people wrote essays, all individuals concerned about what is happening today in the age of AI. *Possible Minds* is not *my* book, rather it is *our* book: Seth Lloyd, Judea Pearl, Stuart Russell, George Dyson, Daniel C. Dennett, Rodney Brooks, Frank Wilczek, Max Tegmark, Jaan Tallinn, Steven Pinker, David Deutsch, Tom Griffiths, Anca Dragan, Chris Anderson, David Kaiser, Neil Gershenfeld, W. Daniel Hillis, Venki Ramakrishnan, Alex "Sandy" Pentland, Hans Ulrich Obrist, Alison Gopnik, Peter Galison, George M. Church, Caroline A. Jones, and Stephen Wolfram.

I see the Possible Minds Project as an ongoing dynamical emergent system, a presentation of the ideas of a community of sophisticated thinkers who are bringing their experience and erudition to bear in challenging the prevailing digital AI narrative as they communicate their thoughts to one another. The aim is to present a mosaic of views that will help make sense out of this rapidly emerging field.

I asked the essayists to consider:

a. The Zen-like poem "Thirteen Ways of Looking at a Blackbird" by Wallace Stevens, which he insisted was "not meant to be a collection of epigrams or of ideas, but of sensations." It is an exercise in "perspectivism," consisting of short, separate sections, each of which mentions blackbirds in some way. The poem is about his own imagination; it concerns what he attends to.
b. The parable of the blind men and an elephant. Like the elephant, AI is too big a topic for any one perspective, never mind the fact that no two people seem to see things the same way.

What do we want the book to do? Stewart Brand has noted that "revisiting pioneer thinking is perpetually useful. And it gives a long perspective that invites thinking in decades and centuries about the subject. All contemporary discussion is bound to age badly and immediately without the longer perspective."

Danny Hillis wants people in AI to realize how they've been programmed by Wiener's book. "You're executing its road map," he says, "and you just don't realize it."

Dan Dennett would like to "let Wiener emerge as the ghost at the banquet. Think of it as a source of hybrid vigor, a source of unsettling ideas to shake up the established mind-set."

Neil Gershenfeld argues that "stealth remedial education for the people running the 'Big Five' would be a great output from the book."

Freeman Dyson, one of the few people alive who knew Wiener, notes that "*The Human Use of Human Beings* is one of the best books ever written. Wiener got almost everything right. I will be interested to see what your bunch of wizards will do with it."

THE EVOLVING AI NARRATIVE

Things have changed—and they remain the same. Now AI is everywhere. We have the Internet. We have our smartphones. The founders of the dominant companies—the companies that hold "the whip that lashes us"—have net worths of $65 billion, $90 billion, $130 billion. High-profile individuals such as Elon Musk, Nick Bostrom, Martin Rees, Eliezer Yudkowsky, and the late Stephen Hawking have issued dire warnings about AI, resulting in the ascendancy of well-funded institutes tasked with promoting "Nice AI." But will we, as a species, be able to control a fully realized, unsupervised, self-improving AI? Wiener's warnings and admonitions in *The Human Use of Human Beings* are now very real, and they need to be looked at anew by researchers at the forefront of the AI revolution. Here is Dyson again:

> Wiener became increasingly disenchanted with the "gadget worshipers" whose corporate selfishness brought "motives to

automatization that go beyond a legitimate curiosity and are sinful in themselves." He knew the danger was not machines becoming more like humans but humans being treated like machines. "The world of the future will be an ever more demanding struggle against the limitations of our intelligence," he warned in *God & Golem, Inc.*, published in 1964, the year of his death, "not a comfortable hammock in which we can lie down to be waited upon by our robot slaves."

It's time to examine the evolving AI narrative by identifying the leading members of that mainstream community along with the dissidents and presenting their counternarratives in their own voices.

The essays that follow thus constitute a much-needed update from the field.

—John Brockman
New York, 2019

Chapter 1

WRONG, BUT MORE RELEVANT THAN EVER

SETH LLOYD

***Seth Lloyd** is a theoretical physicist at MIT, Nam P. Suh Professor in the Department of Mechanical Engineering, and an external professor at the Santa Fe Institute.*

I met **Seth Lloyd** in the late 1980s, when new ways of thinking were everywhere: the importance of biological organizing principles, the computational view of mathematics and physical processes, the emphasis on parallel networks, the importance of nonlinear dynamics, the new understanding of chaos, connectionist ideas, neural networks, and parallel distributive processing. The advances in computation during that period provided us with a new way of thinking about knowledge.

Seth likes to refer to himself as a quantum mechanic. He is internationally known for his work in the field of quantum computation, which attempts to harness the exotic properties of quantum theory, like superposition and entanglement, to solve problems that would take several lifetimes to solve on classical computers.

In the essay that follows, he traces the history of information theory from Norbert Wiener's prophetic insights to the predictions of a technological "singularity" that some would have us believe will supplant the human species. His takeaway on the recent programming method known as deep learning is to call for a more modest set of expectations; he notes that despite AI's enormous advances, robots "still can't tie their own shoes."

It's difficult for me to talk about Seth without referencing his relationship with his friend and professor, the late theoretical physicist Heinz Pagels of Rockefeller University. The graduate student and the professor each had a profound effect on the other's ideas.

In the summer of 1988, I visited Heinz and Seth at the Aspen Center for Physics. Their joint work on the subject of complexity was featured in the current issue of *Scientific American*; they were ebullient. That was just two weeks before Heinz's tragic death in a hiking accident while descending Pyramid Peak with Seth. They were talking about quantum computing.

The Human Use of Human Beings, Norbert Wiener's 1950 popularization of his highly influential book *Cybernetics: or Control and Communication in the Animal and the Machine* (1948), investigates the interplay between human beings and machines in a world in which machines are becoming ever more computationally capable and powerful. It is a remarkably prescient book, and remarkably wrong. Written at the height of the Cold War, it contains a chilling reminder of the dangers of totalitarian organizations and societies, and of the danger to democracy when it tries to combat totalitarianism with totalitarianism's own weapons.

Wiener's *Cybernetics* looked in close scientific detail at the process of control via feedback. ("Cybernetics," from the ancient Greek for "helmsman," is the etymological basis of our word "governor," which is what James Watt called his pathbreaking feedback control device that transformed the use of steam engines.) Because he was immersed in problems of control, Wiener saw the world as a set of complex, interlocking feedback loops, in which sensors, signals, and actuators such as engines interact via an intricate exchange of signals and information. The engineering applications of *Cybernetics* were tremendously influential and effective, giving rise to rockets, robots, automated assembly lines, and a host of precision-engineering techniques—in other words, to the basis of contemporary industrial society.

Wiener had greater ambitions for cybernetic concepts, however, and in *The Human Use of Human Beings* he spells out his thoughts on its application to topics as diverse as Maxwell's Demon, human language, the brain, insect metabolism, the legal system, the role of technological innovation in government, and religion. These broader applications of cybernetics were an almost unequivocal failure. Vigorously hyped from the late 1940s to the early 1960s—to a degree similar to the hype of computer and communication technology that led to the dot-com crash of 2000–2001—cybernetics delivered satellites and telephone switching systems but generated few if any useful developments in social organization and society at large.

Nearly seventy years later, however, *The Human Use of Human Beings* has more to teach us humans than it did the first time around. Perhaps the most remarkable feature of the book is that it introduces a large number of topics concerning human/machine interactions that are still of considerable relevance. Dark in tone, the book makes several predictions about disasters to come in the second half of the 20th century, many of which are almost identical to predictions made today about the second half of the 21st.

For example, Wiener foresaw a moment in the near future of 1950 in which humans would cede control of society to a cybernetic artificial intelligence, which would then proceed to wreak havoc on humankind. The automation of manufacturing, Wiener predicted, would both create large advances in productivity and displace many workers from their jobs —a sequence of events that did indeed come to pass in the ensuing decades. Unless society could find productive occupations for these displaced workers, Wiener warned, revolt would ensue.

But Wiener failed to foresee crucial technological developments. Like pretty much all technologists of the 1950s, he failed to predict the computer revolution. Computers, he thought, would eventually fall in price from hundreds of thousands of (1950s) dollars to tens of thousands; neither he nor his compeers anticipated the tremendous explosion of computer power that would follow the development of the transistor and the integrated circuit. Finally, because of his emphasis on control, Wiener could not foresee a technological world in which innovation and self-organization bubble up from the bottom rather than being imposed from the top.

Focusing on the evils of totalitarianism (political, scientific, and religious), Wiener saw the world in a deeply pessimistic light. His book warned of the catastrophe that awaited us if we didn't mend our ways, fast. The current world of human beings and machines, more than a half century after its publication, is much more complex, richer, and contains a much wider variety of political, social, and scientific systems than he was able to envisage. The warnings of what will happen if we get it wrong, however—for example, control of the entire Internet by a global totalitarian regime—remain as relevant and pressing today as they were in 1950.

WHAT WIENER GOT RIGHT

Wiener's most famous mathematical works focused on problems of signal analysis and the effects of noise. During World War II, he developed techniques for aiming antiaircraft fire by making models that could predict the future trajectory of an airplane by extrapolating from its past behavior. In *Cybernetics* and in *The Human Use of Human Beings*, Wiener notes that this past behavior includes quirks and habits of the human pilot, thus a mechanized device can predict the behavior of humans. Like Alan Turing, whose Turing Test suggested that computing machines could give responses to questions that were indistinguishable from human responses, Wiener was fascinated by the notion of capturing human behavior by mathematical description. In the 1940s, he applied his knowledge of control and feedback loops to neuromuscular feedback in living systems, and was responsible for bringing Warren McCulloch and Walter Pitts to MIT, where they did their pioneering work on artificial neural networks.

Wiener's central insight was that the world should be understood in terms of information. Complex systems, such as organisms, brains, and human societies, consist of interlocking feedback loops in which signals exchanged between subsystems result in complex but stable behavior. When feedback loops break down, the system goes unstable. He constructed a compelling picture of how complex biological systems function, a picture that is by and large universally accepted today.

Wiener's vision of information as the central quantity in governing the behavior of complex systems was remarkable at the time. Nowadays, when cars and refrigerators are jammed with microprocessors and much of human society revolves around computers and cell phones connected by the Internet, it seems prosaic to emphasize the centrality of information, computation, and communication. In Wiener's time, however, the first digital computers had only just come into existence, and the Internet was not even a twinkle in the technologist's eye.

Wiener's powerful conception of not just engineered complex systems but all complex systems as revolving around cycles of signals and computation led to tremendous contributions to the development of complex human-made systems. The methods he and others developed for

the control of missiles, for example, were later put to work in building the Saturn V moon rocket, one of the crowning engineering achievements of the 20th century. In particular, Wiener's applications of cybernetic concepts to the brain and to computerized perception are the direct precursors of today's neural-network-based deep-learning circuits, and of artificial intelligence itself. But current developments in these fields have diverged from his vision, and their future development may well affect the human uses both of human beings and of machines.

WHAT WIENER GOT WRONG

It is exactly in the extension of the cybernetic idea to human beings that Wiener's conceptions missed their target. Setting aside his ruminations on language, law, and human society for the moment, look at a humbler but potentially useful innovation that he thought was imminent in 1950. Wiener notes that prosthetic limbs would be much more effective if their wearers could communicate directly with their prosthetics by their own neural signals, receiving information about pressure and position from the limb and directing its subsequent motion. This turned out to be a much harder problem than Wiener envisaged: Seventy years down the road, prosthetic limbs that incorporate neural feedback are still in the very early stages. Wiener's concept was an excellent one—it's just that the problem of interfacing neural signals with mechanical-electrical devices is hard.

More significantly, Wiener (along with pretty much everyone else in 1950) greatly underappreciated the potential of digital computation. As noted, Wiener's mathematical contributions were to the analysis of signals and noise and his analytic methods apply to continuously varying, or analog, signals. Although he participated in the wartime development of digital computation, he never foresaw the exponential explosion of computing power brought on by the introduction and progressive miniaturization of semiconductor circuits. This is hardly Wiener's fault: The transistor hadn't been invented yet, and the vacuum-tube technology of the digital computers he was familiar with was clunky, unreliable, and unscalable to ever larger devices. In an appendix to the 1948 edition of

Cybernetics, he anticipates chess-playing computers and predicts that they'll be able to look two or three moves ahead. He might have been surprised to learn that within half a century a computer would beat the human world champion at chess.

TECHNOLOGICAL OVERESTIMATION AND THE EXISTENTIAL RISKS OF THE SINGULARITY

When Wiener wrote his books, a significant example of technological overestimation was about to occur. The 1950s saw the first efforts at developing artificial intelligence, by researchers such as Herbert Simon, John McCarthy, and Marvin Minsky, who began to program computers to perform simple tasks and to construct rudimentary robots. The success of these initial efforts inspired Simon to declare that "machines will be capable, within twenty years, of doing any work a man can do." Such predictions turned out to be spectacularly wrong. As they became more powerful, computers got better and better at playing chess because they could systematically generate and evaluate a vast selection of possible future moves. But the majority of predictions of AI, e.g., robotic maids, turned out to be illusory. When Deep Blue beat Garry Kasparov at chess in 1997, the most powerful room-cleaning robot was a Roomba, which moved around vacuuming at random and squeaked when it got caught under the couch.

Technological prediction is particularly chancy, given that technologies progress by a series of refinements, halted by obstacles and overcome by innovation. Many obstacles and some innovations can be anticipated, but more cannot. In my own work with experimentalists on building quantum computers, I typically find that some of the technological steps I expect to be easy turn out to be impossible, whereas some of the tasks I imagine to be impossible turn out to be easy. You don't know until you try.

In the 1950s, partly inspired by conversations with Wiener, John von Neumann introduced the notion of the "technological singularity."

Technologies tend to improve exponentially, doubling in power or sensitivity over some interval of time. (For example, since 1950, computer technologies have been doubling in power roughly every two years, an observation enshrined as Moore's Law.) Von Neumann extrapolated from the observed exponential rate of technological improvement to predict that "technological progress will become incomprehensively rapid and complicated," outstripping human capabilities in the not too distant future. Indeed, if one extrapolates the growth of raw computing power—expressed in terms of bits and bit flips—into the future at its current rate, computers should match human brains sometime in the next two to four decades (depending on how one estimates the information-processing power of human brains).

The failure of the initial overly optimistic predictions of AI dampened talk about the technological singularity for a few decades, but since the 2005 publication of Ray Kurzweil's *The Singularity IS Near,* the idea of technological advance leading to superintelligence is back in force. Some believers, Kurzweil included, regard this singularity as an opportunity: Humans can merge their brains with the superintelligence and thereby live forever. Others, such as Stephen Hawking and Elon Musk, worried that this superintelligence would prove to be malign and regarded it as the greatest existing threat to human civilization. Still others, including some of the contributors to the present volume, think such talk is overblown.

Wiener's lifework and his failure to predict its consequences are intimately bound up in the idea of an impending technological singularity. His work on neuroscience and his initial support of McCulloch and Pitts adumbrated the startlingly effective deep-learning methods of the present day. Over the past decade, and particularly in the last five years, such deep-learning techniques have finally exhibited what Wiener liked to call *Gestalt*—for example, the ability to recognize that a circle is a circle even if when slanted sideways it looks like an ellipse. His work on control, combined with his work on neuromuscular feedback, was significant for the development of robotics and is the inspiration for neural-based human/machine interfaces. His lapses in technological prediction, however, suggest that we should take the notion of a technological singularity with a grain of salt. The general difficulties of technological prediction and the problems specific to the development of

a superintelligence should warn us against overestimating both the power and the efficacy of information processing.

THE ARGUMENTS FOR SINGULARITY SKEPTICISM

No exponential increase lasts forever. An atomic explosion grows exponentially, but only until it runs out of fuel. Similarly, the exponential advances in Moore's Law are starting to run into limits imposed by basic physics. The clock speed of computers maxed out at a few gigahertz a decade and a half ago, simply because the chips were starting to melt. The miniaturization of transistors is already running into quantum-mechanical problems due to tunneling and leakage currents. Eventually, the various exponential improvements in memory and processing driven by Moore's Law will grind to a halt. A few more decades, however, will probably be time enough for the raw information-processing power of computers to match that of brains—at least by the crude measures of number of bits and number of bit-flips per second.

Human brains are intricately constructed, the process of millions of years of natural selection. In Wiener's time, our understanding of the architecture of the brain was rudimentary and simplistic. Since then, increasingly sensitive instrumentation and imaging techniques have shown our brains to be far more varied in structure and complex in function than Wiener could have imagined. I recently asked Tomaso Poggio, one of the pioneers of modern neuroscience, whether he was worried that computers, with their rapidly increasing processing power, would soon emulate the functioning of the human brain. "Not a chance," he replied.

The recent advances in deep learning and neuromorphic computation are very good at reproducing a particular aspect of human intelligence focused on the operation of the brain's cortex, where patterns are processed and recognized. These advances have enabled a computer to beat the world champion not just of chess but of Go, an impressive feat, but they're far short of enabling a computerized robot to tidy a room. (In

fact, robots with anything approaching human capability in a broad range of flexible movements are still far away—search “robots falling down.” Robots are good at making precision welds on assembly lines, but they still can’t tie their own shoes.)

Raw information-processing power does not mean sophisticated information-processing power. While computer power has advanced exponentially, the programs by which computers operate have often failed to advance at all. One of the primary responses of software companies to increased processing power is to add “useful” features, which often make the software harder to use. Microsoft Word reached its apex in 1995 and has been slowly sinking under the weight of added features ever since. Once Moore’s Law starts slowing down, software developers will be confronted with hard choices between efficiency, speed, and functionality.

A major fear of the singulariteers is that as computers become more involved in designing their own software they’ll rapidly bootstrap themselves into achieving superhuman computational ability. But the evidence of machine learning points in the opposite direction. As machines become more powerful and capable of learning, they learn more and more as human beings do—from multiple examples, often under the supervision of human and machine teachers. Education is as hard and slow for computers as it is for teenagers. Consequently, systems based on deep learning are becoming more rather than less human. The skills they bring to learning are not “better than” but “complementary to” human learning: Computer learning systems can identify patterns that humans cannot—and vice versa. The world’s best chess players are neither computers nor humans but humans working together with computers. Cyberspace is indeed inhabited by harmful programs, but these primarily take the form of malware—viruses notable for their malign mindlessness, not for their superintelligence.

WHITHER WIENER

Wiener noted that exponential technological progress is a relatively modern phenomenon and not all of it is good. He regarded atomic

weapons and the development of missiles with nuclear warheads as a recipe for the suicide of the human species. He compared the headlong exploitation of the planet's resources with the Mad Tea Party of *Alice in Wonderland*: Having laid waste to one local environment, we make progress simply by moving on to lay waste to the next. Wiener's optimism about the development of computers and neuromechanical systems was tempered by his pessimism about their exploitation by authoritarian governments, such as the Soviet Union, and the tendency for democracies, such as the United States, to become more authoritarian themselves in confronting the threat of authoritarianism.

What would Wiener think of the current human use of human beings? He would be amazed by the power of computers and the Internet. He would be happy that the early neural nets in which he played a role have spawned powerful deep-learning systems that exhibit the perceptual ability he demanded of them—although he might not be impressed that one of the most prominent examples of such computerized *Gestalt* is the ability to recognize photos of kittens on the World Wide Web. Rather than regarding machine intelligence as a threat, I suspect he would regard it as a phenomenon in its own right, different from and co-evolving with our own human intelligence.

Unsurprised by global warming—the Mad Tea Party of our era—Wiener would applaud the exponential improvement in alternative-energy technologies and would apply his cybernetic expertise to developing the intricate set of feedback loops needed to incorporate such technologies into the coming smart electrical grid. Nonetheless, recognizing that the solution to the problem of climate change is at least as much political as it is technological, he would undoubtedly be pessimistic about our chances of solving this civilization-threatening problem in time. Wiener hated hucksters—political hucksters most of all—but he acknowledged that hucksters would always be with us.

It's easy to forget just how scary Wiener's world was. The United States and the Soviet Union were in a full-out arms race, building hydrogen bombs mounted on nuclear warheads carried by intercontinental ballistic missiles guided by navigation systems to which Wiener himself—to his dismay—had contributed. I was four years old when Wiener died. In 1964, my nursery school class was practicing duck and cover under our desks to prepare for a nuclear attack. Given the

human use of human beings in his own day, if he could see our current state, Wiener's first response would be to be relieved that we are still alive.

Chapter 2

THE LIMITATIONS OF OPAQUE LEARNING MACHINES

JUDEA PEARL

Judea Pearl *is a professor of computer science and director of the Cognitive Systems Laboratory at UCLA. His most recent book, co-authored with Dana Mackenzie, is* The Book of Why: The New Science of Cause and Effect.

In the 1980s, **Judea Pearl** introduced a new approach to artificial intelligence called Bayesian networks. This probability-based model of machine reasoning enabled machines to function—in a complex and uncertain world—as "evidence engines," continuously revising their beliefs in light of new evidence.

Within a few years, Judea's Bayesian networks had completely overshadowed the previous rule-based approaches to artificial intelligence. The advent of deep learning—in which computers, in effect, teach themselves to be smarter by observing tons of data—has given him pause, because this method lacks transparency.

While recognizing the impressive achievements in deep learning by colleagues such as Michael I. Jordan and Geoffrey Hinton, he feels uncomfortable with this kind of opacity. He set out to understand the theoretical limitations of deep-learning systems and points out that basic barriers exist that will prevent them from achieving a human kind of intelligence, no matter what we do. Leveraging the computational benefits of Bayesian networks, Judea realized that the combination of simple graphical models and data could also be used to represent and infer cause-effect relationships. The significance of this discovery far transcends its roots in artificial intelligence. His latest book explains causal thinking to the general public; you might say it is a primer on how to think even though human.

Judea's principled, mathematical approach to causality is a profound contribution to the realm of ideas. It has already benefited virtually every field of inquiry, especially the data-intensive health and social sciences.